AL-TR-1992-0044

ARMSTRONG LABORATORY

EVALUATING TRAINING AND EDUCATIONAL PROGRAMS: A REVIEW OF THE LITERATURE

Joseph S. Mattoon

HUMAN RESOURCES DIRECTORATE
AIRCREW TRAINING RESEARCH DIVISION
Williams Air Force Base, AZ 85240-6457

October 1992

Final Technical Report for Period November 1988 – November 1991

Approved for public release; distribution is unlimited.

AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5000

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

JOSEPH S. MATTOON
Project Scientist

DEE H. ANDREWS, Technical Director
Aircrew Training Research Division

LYNN A. CARROLL, Colonel, USAF
Chief, Aircrew Training Research Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1992	Final – November 1988 – November 1991

4. TITLE AND SUBTITLE
Evaluating Training and Educational Programs: A Review of the Literature

5. FUNDING NUMBERS
PE – 62205F
PR – 1123
TA – 31
WU – 04

6. AUTHOR(S)
Joseph S. Mattoon

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Armstrong Laboratory
Human Resources Directorate
Aircrew Training Research Division
Williams Air Force Base, AZ 85240-6457

8. PERFORMING ORGANIZATION REPORT NUMBER
AL-TR-1992-0044

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

A review of the literature was conducted to describe the field of program evaluation, to identify useful sources of information on this topic, and to outline concepts and methodologies that have been proposed for planning and conducting evaluations. The failures of early program evaluations led to new organizations that attempted to clarify the field and improve the effectiveness of evaluators. Several sources of information on program evaluation are recommended reading for those who are preparing to conduct evaluation activities. Since the 1960s, evaluation researchers have expanded the concept and definition of program evaluation and have assembled a substantial number of new methods, tools, and strategies that are valuable to practitioners. The most promising predictor of success of program evaluation is proper planning. Evaluators should work closely with program decision makers and sponsors from the planning stage throughout the life of the program. Formative evaluation should be conducted throughout the life of the program. Summative measures should be conducted on programs only after they have become fully operational.

14. SUBJECT TERMS
Contract monitoring
Evaluation
Program assessment
Program evaluation
Program management
Training evaluation
Training programs

15. NUMBER OF PAGES
56

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

CONTENTS

PREFACE

Evaluation is an essential component of all training programs, but the complexity of program environments and the magnitude of evaluation tasks make it a very challenging endeavor. Historically, program evaluators have had to deal with a multitude of problems that range from the lack of agreement on the purpose of evaluating programs to disagreements on appropriate performance measurement techniques. This report represents a summary of the literature on program evaluation that was published from the mid 1960s to the present. Its purpose is to (a) clarify the purpose, goals, and objectives of program evaluation, (b) identify sources of information useful to evaluators, and (c) outline strategies, methods, and tools that have been recommended by experts in the field. The work was conducted under Work Unit 1123-31-04, C-130 Aircrew Training Systems Research Program. The work unit monitor was Dr. Robert T. Nullmeyer, and the principal investigator was Mr. Joseph S. Mattoon.

EVALUATING TRAINING AND EDUCATIONAL PROGRAMS: A REVIEW OF THE LITERATURE

INTRODUCTION

A review of the literature was conducted for the purpose of describing the field of program evaluation, identifying useful sources of information on this topic, and outlining concepts and methodologies that have been proposed for planning and conducting evaluations. This report focuses on education and training programs, but many of the problems and solutions discussed are generalizable to other program environments. Evaluation connotes different concepts, purposes, or activities depending on one's professional background. For example, an electrical engineer may think of evaluation as a highly structured set of technical procedures for testing instruments, while a public administrator may consider evaluation to be the process of using intuitive judgment to analyze the interactions of people. Program evaluation refers to the examinations of both people and resources organized for the purpose of attaining particular goals. Generally speaking, the purpose of program evaluation is to verify the effectiveness and efficiency of a program and to generate information that will facilitate its improvement on attaining organizational goals. The information in this literature review is intended for practitioners, researchers, and those who are involved in the commission of program evaluations. It may be especially interesting and useful for anyone who is planning to participate in an evaluation of a program and who has no prior evaluation experience.

This report begins by introducing program evaluation as an important but problematic field. The second part of the review describes the characteristics of different types of publications containing information on evaluation research and practice. The major portion of this report describes the development of program evaluation, five general problem areas that effect the attainment of evaluation goals, and the recommendations of evaluation researchers for avoiding previously encountered problems in planning, conducting, and using the results of program evaluations. The conclusion summarizes the most important issues found within the literature and the lessons learned in program evaluation.

Public education programs share some key attributes with military and other large-scale training programs because they are all instructional systems. Each involves similar activities: (a) defining program goals in terms of the learning benefits of their recipients (students or trainees); (b) planning and assembling curriculum and the design and development of courses and procedures for instructional management; and (c) evaluation of the program to determine the degree that recipients gain the desired benefits and the degree that the program operates efficiently within its budget and available resources. Educational programs have undergone many types of evaluation in the last 20 years. The lessons learned in the effort to evaluate these programs have resulted in a good deal of debate over the philosophy of program evaluation and appropriate tools and methodology for practitioners.

Historically, crises have been the prerequisite of wisdom and the precursor to change in the field of program evaluation. The failures of educational and social program evaluations conducted in the early 1960s led to major research efforts to formulate better strategies for evaluating large programs (Stufflebeam, Foley, Gephart, Guba, Hammond, Merriman, & Provus, 1971). During this period the U.S. Government began to allocate a good deal of capital for the purpose of evaluating new public education programs (Stufflebeam & Webster, 1981). A plan for program evaluation became a mandatory component of each new program proposed, and submitting evaluation results became mandatory for the continuation of program funding. However, early efforts to meet these requirements proved ineffective. This poor performance produced a crisis that forced evaluators to devise new theories and methodologies in an attempt to meet government mandates. The crisis eventually resulted in the launching of new professional organizations such as the Phi Delta Kappa National Study Committee on Evaluation (PDK Committee), the Joint Committee on Standards for Educational Evaluation, the Evaluation Research Society (ERS); and more recently, the American Evaluation Association which is a joint organization of Evaluation Network and ERS. A number of professional journals (e.g., Evaluation Practice and Evaluation Review) became available and were published to disseminate information and debate new evaluation philosophies and methodologies.

The push to evaluate new programs in the 1960s and the subsequent failure of evaluators to meet the needs of program sponsors, personnel, and recipients revealed serious deficits in the field of evaluation. Program evaluators discovered that methods which have been developed for evaluating individual products and conducting laboratory research failed when applied to large, complex educational programs (Stufflebeam, et al., 1971). The inability of evaluators to empirically demonstrate a program's success produced two major problems: (a) There was no way to justify funding to continue a program. (b) Program administrators had no guidance for initiating improvements.

In 1966 the Phi Delta Kappa Research Advisory Board recommended the establishment of a special committee to critique evaluation theories and models and to develop an effective methodology for program evaluation. The PDK Committee was established by recruiting members from the Evaluation Center at Ohio State University, the Research and Development Center on Evaluation at the University of Los Angeles, and the EPIC Evaluation Center in Tucson, Arizona. The PDK Committee identified a number of serious problems in the field, and their findings appear to have been a pivotal point in the generation of new philosophy and methodology for program evaluation. The PDK Committee's report represents an extensive examination of the field. The findings and recommendations of their report are generalizable to current programs. This review focuses on the major issues raised in the PDK Committee's report, Educational Evaluation and Decision Making (Stufflebeam, et al., 1971), combined with information excerpted from more recent publications.

SOURCES OF INFORMATION

The first step in identifying useful sources of information on program evaluation was to determine what topics, issues, and information would be most useful to evaluators. Scriven (1986) explains that the descriptions of program evaluation vary greatly because it is an extremely broad field that covers every activity from the evaluation of small educational projects within a single school district to ongoing evaluations of very large programs such as Medicare. A cursory review of the literature was conducted to identify major aspects of program evaluation and was followed by a more in-depth review of particular articles, reports, and books.

The PDK Committee identified four general approaches used to evaluate new programs: (a) Objectives-oriented evaluation defines educational objectives and assesses the degree to which program recipients have attained them. (b) Nationally standardized tests have been developed to reflect the content of new educational curricula and assess its effect on student abilities on a wide scale. (c) The professional judgment method provides ratings for proposals and monitors the progress of contractors. (d) Field experiments are conducted within the environments of new programs to assess their effect on program recipients. Chelimsky (1985) defines "program evaluation" as "the application of systematic research methods to the assessment of program design, implementation, and effectiveness" (p. 488). This definition and others outlined by the PDK Committee were used to select the type of literature most appropriate for this review.

Program evaluation has been derived in an eclectic fashion through a synthesis of theories and methodologies taken from a number of disciplines such as psychology, statistical analysis, economics, and political science. These disciplines share some characteristics, but are fundamentally different in several ways. Evaluations vary as a function of evaluators' experience within different disciplines, different program contexts, and different needs of those who commission evaluations (Jemelka & Borich, 1979). Because program evaluation covers such a broad range of human endeavor, useful information may be found in several different types of literature depending on the needs of the reader. However, this diversity within the field has led to confusion and misunderstanding of the purpose and methodologies of program evaluation.

Three types of literature, published from the early 1970s to the present were reviewed--evaluation reports, professional journals, and books. Evaluation reports originated from several agencies within the Federal Government and government contractors. Professional journals featured articles with a strong emphasis on either evaluation methods and theory or results of specific evaluations. After a preliminary examination of each type of publication, a number of individual evaluation reports, four periodicals and 14 books were selected for a more thorough review. The kind of information sought was that which revealed the philosophy and methodology of program evaluation, the most common problems encountered in the practice of program evaluation, and proposed solutions to program evaluation problems.

A search of the Monthly Catalogue of Government Publications (Superintendent of Documents, 1978-1989) produced 30 documents consisting of single evaluation reports and volumes of reports. Most of these were published by the Air Force Human Resources Laboratory (AFHRL), Naval Training Systems Center (NTSC), U.S. Army Research Institute for Behavioral and Social Sciences (ARI), and the General Accounting Office (GAO). Reports by organizations within the military services tended to focus on evaluations of training devices such as the F-111 Converter/Flight Control System Simulator (Ciechinelli, Harmon, Keller, 1982) and assessment techniques such as peer performance assessment (ARI, 1978). GAO's reports focus on monetary expenditures and accountability rather than effectiveness of devices or methods. For example, one of the GAO reports that was reviewed describes costs associated with the acquisition of two computer systems for U.S. air bases (Comptroller General of the U.S., 1979). The main function of this report was to recommend the acceptance or rejection of a program proposal rather than discussing or advocating evaluation methods. A volume of evaluation reports, published quarterly by Health and Human Services (HHS), contained abstracts of hundreds of evaluation and research studies conducted on medical innovations, products, and programs within HHS (HHS Evaluation and Documentation Center, 1985). It was organized by issues such as "Research and Evaluation Performance" and "Research and Evaluation Planning and Administration." Most reports published by HHS concentrated on evaluation methods developed for specific health practices.

After reviewing and summarizing government evaluation reports, it was concluded that their main purpose is to describe the strengths and weaknesses of products, methods, and proposals. Because these reports seldom articulated the theories and methodologies for conducting program evaluations, further in-depth review of this literature was not conducted.

Surveys by Turpin, Smith, and Darcy (1987) and by Shadish and Reichardt (1988) have identified a number of journals that regularly publish articles focused on program evaluation. Turpin et al. (1987) conducted an extensive survey of 92 journal editors representing 80 journals. Editors were asked to describe characteristics such as the "types of articles" published and "emerging trends in evaluation research." Their survey identified 33 journals that publish evaluation research. Shadish and Reichardt (1988) used a different approach to select journals. They reviewed 469 articles reprinted in the Evaluation Studies Review Annual (ESRA) representing 123 journals, eight of which published more than 10 articles in past ESRA volumes. These two surveys were used as a tool for selecting literature.

Thirteen periodicals initially selected for review in the present investigation were: Evaluation Review (formerly Evaluation Quarterly), International Journal of Educational Research (formerly Evaluation of Education), Performance Improvement Quarterly, Evaluation Practice (formerly Evaluation News), Evaluation and Program Planning, Applied Psychology, Journal of Human Resources, Psychological Bulletin, Journal of Instructional Development, Evaluation Comment, Educational Evaluation and Policy Analysis, Harvard Educational Review, and Journal of Policy Analysis and Management. After reviewing the most recent

issues of each periodical, three journals and ESRA were selected for a more thorough examination. One of these periodicals is a bimonthly journal, Evaluation Review; two are quarterly journals, Evaluation and Program Planning and Evaluation Practice, and ESRA is a review of publications that reprints key articles published within a given year from different journals. Articles from the four selected periodicals were reviewed beginning with the first issue published.

Fourteen books were also chosen for review based on the frequency and content of authors' publications in professional journals and on the frequency that their works were cited by other authors in the field of program evaluation. About half of the content in this report is based on information taken from books and the other half from journal articles.

There are two basic types of publications on program evaluation, those that describe the results of case studies and those that describe theories and methodologies of evaluation. The majority of evaluation reports describe case studies, while journal articles and books tend to focus on theory and methodology. The issues covered in journals and books are often referred to as "evaluation research." Nagel (1985) describes evaluation research as ". . . the development of general principles of evaluation" (p. 61). Although journals typically show some degree of specialization (e.g., public education), most contain articles on a variety of topics. For example, Evaluation Review provides many reports on particular case studies but also addresses broader issues within evaluation research. Journals also feature comparisons of conflicting views and philosophies such as *qualitative versus quantitative methods*. Edited books offer comprehensive coverage of evaluation philosophy and methodology, but they tend to be dominated by particular interpretations of evaluation that may be contrary to the views of other authors.

If one wishes to inquire about the findings of evaluations for specific projects or products, government documents should prove most useful. However, if more general information is needed, journals such as the three chosen for this review would be more appropriate. Finally, for detailed coverage of a particular evaluation model, methodology, or approach to program evaluation, books provide the most comprehensive information.

THE DEVELOPMENT OF PROGRAM EVALUATION

A great enthusiasm for developing new educational programs and improving existing programs was spurred on by several major events in the mid- and late twentieth century: (a) the launching of Sputnik I by the Soviet Union in 1957 (Cronbach, Ambron, Dornbusch, Hess, Hornick, Phillips, Walker, & Weiner, 1980); (b) the adoption of the Planning Programming and Budgeting System (PPBS) during the period of President Johnson's "Great Society" programs (Thompson, 1982); (c) the Elementary and Secondary Education Act (ESEA) of 1965; and (d) the accountability movement beginning in the early 1970s (Rutman, 1980). During the 1960s and 1970s, changes in attitudes toward government spending brought about mandatory requirements that linked program evaluation to budgetary

decisions. Congressional actions such as the Congressional Budget and Improvement Control Act of 1974 demonstrate a strong government initiative for evaluating federal programs.

Economic constraints have produced pressures for military training. Bruce (1989) states that the "Department of Defense Directive No. 1430.13, Training Simulations and Devices (Office of Secretary of Defense, 22 August 1986) specifies that training effectiveness evaluations are to be conducted to ensure that devices meet training requirements and effectiveness levels" (p. 1). In U.S. Air Force (USAF) training programs, the aircraft is still the primary medium for training (Bruce, Killion, & Rockway, 1989). Because of the great costs associated with flying hours, it is extremely important to achieve training effectiveness. Most flying skills taught in ground-based training programs are developed using simulators. Although more economical than aircraft for training, flight simulators are expensive to build, operate, and maintain.

In a report on B-52 and KC-135 mission qualification and continuation training, Bruce, Killion, and Rockway (1989) concluded that operational units are encouraged to ". . . optimize the allocation of limited training resources to an expanding number of training requirements" (p. 28). They also point out that a need exists to optimize the use of individual media and download some of the aircraft-based training to alternative devices (simulators). Finally, a need for more efficient and accurate skill assessment is needed because many proficiency assessment methods are event-based and measure how much training is done rather than the abilities of training recipients. These problems require a thorough analysis of USAF training programs which may be accomplished through program evaluation.

The enormous responsibility that is placed upon evaluators is apparent when one considers the consequences of continuing or discontinuing funding for state- and even nation-wide programs. Many of these large programs provide jobs for hundreds of people and involve millions of dollars. Thompson (1982) writes that ". . . offices of evaluation [have] . . . guided the spending of hundreds of billions of dollars." With so much at stake, there is a great deal of pressure placed on evaluators by program sponsors, staff members, and other decision makers, and this is probably why many individuals who possess the skills to evaluate programs are choosing more comfortable professions today (Cook & Leviton, 1983).

The passage of ESEA in 1965 brought about mandatory requirements for evaluating new educational programs and much concern for effective evaluation methodology. Millions of dollars were allocated for evaluating new programs in the 1960s, but the evaluations often yielded inconclusive or unusable information. The initiative for improving education and the subsequent failure of many evaluations forced evaluators to invent new concepts and techniques in the attempt to fulfill evaluation mandates. The development of new program evaluation methodologies and efforts to adopt and adapt methodologies from other disciplines began in the late 1960s. These methodologies included the Delphi technique, item sampling, criterion referenced tests, Bayesian statistics,

operations research, systems analysis, Program Evaluation Reporting Technique (PERT), Stake's (1967) Countenance Model for Educational Evaluation, Stufflebeam's CIPP Evaluation Model (Stufflebeam & Shinkfield, 1985), and Scriven's (1967) Consumer-Oriented approach to evaluation. The appearance of these new methods and models represent a turning point in the field of program evaluation.

Cronbach et al. (1980) describes some aspects of the ESEA's specifications for program evaluation that produced a good deal of frustration among school officials and professional evaluators. Two major problems were that guidelines for conducting evaluation and reporting results were too broadly defined to clearly describe what was required, and the field of program evaluation was not sufficiently mature to produce the type of data that was needed to improve programs or even to justify their existence. For example, ESEA required that all proposals for new programs requesting funding through Title I include a description of how funds would be disseminated, the manner in which the program would improve the educational attainment of deprived children, and a summary report by the state agency containing data based on "appropriate objective measures of educational achievement." One problem was that these elements were so broadly stated that almost any procedure could be perceived by evaluators as acceptable. A second problem was that early evaluation techniques were incapable of producing substantive data, and this problem became apparent within a year after the passing of ESEA. Since no specifications were made on exactly what types of data were appropriate, and evaluators did not know how to collect and analyze program data in a way that would determine the effectiveness of new programs, the situation soon became critical.

Cronbach et al. (1980) explain that one of the reasons that ESEA guidelines for evaluation were imprecisely stated was that school officials influenced the design of the act's specifications for evaluation. In the attempt to protect Title I dollars from being wasted on ineffective programs, Senator Robert Kennedy proposed some tough requirements for evaluation of Title I programs, but state and local school officials objected to this tight, Federal control of educational funding. They complained that "for the Federal Government to specify educational objectives and measurement procedures was intrusive and inappropriate." Guideline writers in the Office of Education feared that the bill would not pass because of this outcry, so they produced a "least-restrictive" interpretation of evaluation. Consequently, local schools were "free to report any kind of 'objective' data regarding 'needs' of their own choosing" (p. 33).

ESEA promised to improve educational programs through closer examination of program proposals. It seems that the basic intent of ESEA was sound, but the guidelines were too general or too loosely-stated, and the tools and methods for producing valid, substantive evaluation data were simply not available in the 1960s. A manual published by the U.S. Office of Education on Title III of ESEA mandates that program proposals must include a description of the *procedures, methods, and instruments* to be used to evaluate the degree to which objectives are achieved. Unfortunately, evaluators had not yet developed instruments that were sensitive enough to accomplish such a task. The result

was that many program evaluations that were conducted in compliance with ESEA guidelines produced inaccurate or useless findings. Some even damaged efforts toward program improvement (Roberts-Grey, Buller, & Sparkman, 1987).

There are some important lessons to learn from the failure of ESEA to produce adequate program evaluations. First, it is apparent that guidelines should clearly describe program evaluation procedures and expected outcomes of the evaluation, the type of data to be collected, and the methods and instruments that will render the data useful in describing the program and its effects on recipients. Such guidelines would probably increase the quality of evaluations and also function as a planning tool for program development and implementation. Current literature on program planning and evaluation consistently emphasizes the need for establishing clearly stated program goals and objectives as a joint effort by those who have a stake in the target program. If evaluation plans are developed by only one of these parties, efforts to collect and analyze data will likely be hindered or even blocked by those who disagree with the methodology or feel threatened by the evaluation. In this view, program goals and objectives should be drawn up through a cooperative effort of program administrators, funding agencies, and evaluators prior to development of the program.

Proposals for program evaluation should state a purpose for the evaluation, evaluation procedures, and how results will be used. In the 1960s program planners attempted to propose, implement, and evaluate programs as specified by the ESEA. However, since the ESEA failed to identify specific goals for evaluations and methods for data collection and analyses, evaluations of educational programs produced neither valid evidence of programs' worth or information that assisted in improving programs.

There is evidence that program sponsors did not have much faith in using the results of evaluations to improve programs during the 1960s. For example, a statement from the Citizens' Committee for Children of New York, Inc., in 1967 (cited in Stufflebeam et al., 1971), that was brought before the Subcommittee on ESEA of the Education and Labor Committee, indicated that the educational projects in New York City were recycled in 1967 before the results of the 1966 evaluations were even reported. If plans to rebuild, revise, or discontinue a program are made without considering evaluation results, there is little need to expend the funds and personnel required to conduct program evaluation.

Eight Symptoms of Illness of Program Evaluation

The PDK Committee identified eight general "symptoms of evaluation's illness" that have contributed to failures and lack of enthusiasm for program evaluation (Stufflebeam, et al., 1971, p. 4-9). Recently published work in evaluation research indicates that many of these symptoms persist in training and education programs today. Each of these symptoms are described below. Those symptoms that are most relevant to program evaluation today are supplemented with information from more recent literature.

The Avoidance Symptom

From the level of the U.S. Office of Education (USOE) down to the local school district, the PDK Committee judged the evaluation budget and staff to be too small or inadequate to effectively evaluate programs. For example, it reported that ". . . the Bureau of Elementary and Secondary Education in the USOE, which spends over half the total Federal money available for education (almost $4 billion annually), during fiscal year 1968 budgeted $1,157,000 for various evaluation activities. However, only $30,000 was earmarked for new evaluations . . ." (p. 5).

The Anxiety Symptom

Evaluators experience a good deal of anxiety arising from multiple causes. The evaluator knows that many personnel within a program's staff or management team often share the negative concept that evaluations are designed to pass judgments on their performance. Since evaluation requires the cooperation of these individuals, the evaluator is in a difficult position from the start. Also, the evaluator is aware that the process of evaluation is often cursory, inadequate, and subject to error. Finally, it is well known among evaluators and program managers alike that many evaluative methods and program elements are so poorly understood that an evaluation can yield meaningless data.

The Skepticism Symptom

Despite the need for evaluation and the desirable consequences of conducting evaluations to improve programs, many argue that planning for evaluation is a useless endeavor. This problem is aggravated by the lack of agreement among expert evaluators on the proper approach and methodology. If professional evaluators cannot come to an agreement on what should be done, how can one expect the staff and sponsors of programs to have faith in proposed evaluation designs?

The record of the successive failure of program evaluations combined with the disruption they have produced within programs due to anxiety and skepticism are quite enough to cause both program staff and sponsors to do all they can to avoid evaluation. Hays (1987) states that "evaluation is greeted not with open minds but with resistance, suspicion, and mistrust." She provides three reasons: (a) Evaluators and program staff are uncertain as to whether ". . . evaluation is reliable when applied to complex, irregular, and nonroutinized technologies." (b) The program staff and evaluators are acutely aware of the politically charged environment during evaluation. (c) There are almost always negative attitudes expressed about the purpose of the evaluation in terms of program improvement, accountability, and major decisions that will strongly affect program personnel.

Hays stresses the need for ". . . dialogue between managers and evaluators based on mutual respect for what each can contribute to the other's knowledge about the program and its evaluation" (p. 31). This position is strongly supported by evaluators and evaluation researchers.

Rutman (1980) points out that program managers are understandably nervous about evaluations. Evaluation conclusions have often been used to curtail or even eliminate programs or hold management accountable for solving problems and achieving goals that are not appropriate or are inconsistent with stated goals. The consequences of conducting evaluations in environments where skepticism and anxiety are a way of life are serious. Program personnel are likely to be uncooperative or will attempt to provide only data that will lead to favorable conclusions (Nay & Kay, 1982).

The Lack-of-Guidelines Symptom

Guidelines produced to facilitate compliance to legislative requirements for evaluations are subject to wide interpretation and do not state evaluation processes in operational terms. The inability of the very agencies that require evaluation to provide adequate guidelines is serious.

The Immobilization Symptom

Schools have not responded to legislation in terms of providing personnel and funds for systematic evaluation despite Federal requirements (i.e., Title I and Title III of ESEA), and there are no good models to represent what schools do to comply with these requirements.

The "lack-of-guidelines symptom" appears to be a major contributor to the failure of institutions to comply with Federal requirements for conducting evaluations. Since the publication of the PDK Committee's report, a number of standards for program evaluation have been developed. They are based on concepts and definitions of program evaluation that describe the role of evaluators, the major goals and functions of program evaluation, and the task involved in conducting an evaluation. The Joint Committee on Standards for Program Evaluation (1981) developed 30 such standards over a 5-year period. The full description of these standards can be found in the volume, Standards for Evaluations of Educational Programs, Projects, and Materials. The ERS Standards Committee (1982) developed 55 evaluation standards. Their descriptions are present in Standards for Evaluation Practice: New Directions for Program Evaluation, No. 15. Cronbach and his associates (1980) have produced what they call "Our Ninety-Five Theses"; a list of do's, don't's, standards, problems, and solutions in program evaluation. Their work describes the overall process of program evaluation plus a number of essential evaluation tasks such as setting goals for proposed evaluations and planning evaluation procedures.

In a report to Congress in response to the Educational Amendments of 1978, a group of evaluators from Northwestern University recommended that evaluation standards be recognized because they are useful for clarifying what is meant by the quality of evaluation, and they inform the public of what can be expected of evaluations (Boruch, Cordray, Pion, & Leviton, 1983).

The Misadvice Symptom

An analysis of 21 Title III proposals revealed that evaluation designs are often left out of proposals. Those proposals that did contain evaluation designs had serious deficiencies. The conclusion is that experts in the field have been unable to design evaluations that produce results which can be translated into useful advice for program staff and management. Useful advice refers to information that guides actions which lead to the improvement in programs. Misadvice refers to recommendations that accomplish nothing or hurt programs.

The No-Significant-Difference Symptom

There is a very high frequency of no-difference findings in educational evaluation studies that compare two or more alternative instructional methods. When program evaluation fails to detect effects that are clearly observable, it is necessary to question evaluation methodology.

The Missing-Elements Symptom

The field of evaluation has failed to make significant forward strides. There is a lack of (a) adequate theory; (b) specification for useful evaluative information; (c) appropriate instruments and designs; (d) mechanism for organizing, processing, and reporting evaluative information; and (e) trained personnel.

A great concern over the methodological weakness of program evaluation has been voiced by professionals who are familiar with measurement and human performance within organizations. This inadequacy validates program managers' fears about the accuracy of information and conclusions presented in evaluations. In speaking of the methodology used for evaluating human services and education, Mann (cited in Rutman, 1980) reports that there are doubts as to whether conclusions should ever attempt to be drawn from evaluations. Rutman (1980) points out that poor methodology will not necessarily cause program managers to discard evaluation findings. The greater danger is that erroneous findings derived from poor methods will be used, thus damaging rather than enhancing program performance.

A report by the U.S. General Accounting Office (GAO) paints a bleak picture of the state of program evaluation in the Government today. The report describes program evaluation and program data collection procedures as being

in a depleted state in executive agencies (U.S. General Accounting Office, 1988). The GAO also reports that the number of in-house professionals has been reduced and the pool of professional contractors that do evaluation work is decreasing. Cook and Leviton (1983) state that evaluation is not growing at the rate it did in the 1960s and 1970s and that the field may be declining both in funding support for activities and in the number of trained evaluators; there has been a shift from externally funded evaluations to internal evaluations conducted by special branches of the Government (House, 1986).

General Conclusions about Symptoms of Illness

The PDK Committee indicates that "little effort has been expended to overcome the eight symptoms of illness" and emphasizes the seriousness of the problem:

> Even the best evaluators can function only in terms of extant theory and the available concepts, design, tools, instruments, and training. Any professional area that is so much avoided, produces so many anxieties, immobilizes the very people who might want to avail themselves of it, is so widely regarded with skepticism, is incapable of operational definition even by its most trained advocates (who in fact render bad advice to the practitioners who consult them), is ineffective in answering reasonable and important questions, and has made little apparent effort to isolate and ameliorate its most serious lacks is indeed on the critical list (p. 9).

If such criticism was the product of a single author's opinion, it would carry much less weight. Because the PDK Committee represents professional evaluators who were assembled for the sole purpose of reporting the state of program evaluation, and because almost all of these problems have been verified by more recent evaluation literature, the criticism should be taken seriously. It is not surprising that program evaluation is in such a poor state if one considers its most serious problems together: (a) Substantial gaps still exist in the theory and methodology of program evaluation. (b) The Government is spending less money on evaluation. (c) The responsibility of conducting program evaluations has shifted to in-house personnel. (d) In-house evaluators are on the decline.

Many of the problems of program evaluation in the late 1960s are still present and continue to plague evaluators and program planners today. Therefore, it is worthwhile to elaborate on each problem area and describe some of the general recommendations that have been proposed.

Major Problem Areas in Program Evaluation

The PDK Committee's report describes five major problem areas that have contributed to the eight symptoms of illness described above (Stufflebeam et al., 1971). They are: "(a) definition [of evaluation] (p. 9), (b) decision making

[methodology] (p. 16), (c) [establishing] values and criteria (p. 18), (d) [dealing with different] administrative levels (p. 19), and (e) problems with the [evaluation] research model" (p. 22). A description of each problem area is followed by recommendations based on the conclusions of the PDK Committee and on information from more recent publications. Recommendations sometimes represent contradicting views, an indication that a widely accepted paradigm for program evaluation does not yet exist.

Early Concepts of Program Evaluation

The PDK Committee outlines three early descriptions of program evaluation produced during the early 1960s. Each type of description reveals a particular misconception or incompleteness. The identification of these shortcomings has helped evaluation researchers develop a more definitive understanding of the field and has facilitated the development of several types of evaluation activities that help evaluators plan and conduct more effective program evaluation.

Thorndike and Hagen (1961) describe evaluation as more inclusive than measurement but that measurement techniques provide the foundation for sound evaluation. Ebel (1965) describes evaluation as a judgment of merit based on measurements. There are some distinct drawbacks to these definitions: (a) They limit the view of evaluation to that of an instrument. (b) They tend to obscure the fact that value judgments are always involved in an evaluation. (c) They limit evaluation to the theory and practice of measurement. Many important aspects of a program cannot be measured by conventional instruments that rely on quantitative data. Quantitative analysis is appropriate when applied to certain data sources, but it will lead to erroneous conclusions when data cannot be accurately expressed in numerical form.

A second description is based on determining the congruence among student performance and objectives. This idea originated from Tyler's (1967) work in the late 1940s and 1950s on the Eight-Year Study at Ohio State University. The work of Furst, a student of Tyler's, demonstrated the strong focus of this approach on student performance. Furst (cited in Stufflebeam et al., 1971) defines four evaluation tasks that are based on the goal of establishing congruence between objectives and student performance.

1. To determine the objectives which the course or program should seek to attain.

2. To select learning experiences which will help to bring about the attainment of these objectives.

3. To organize these learning experiences so as to provide continuity and sequence for the student and to help him integrate what might otherwise appear as isolated experiences.

4. To determine the extent to which the objectives are being attained (p. 12).

Furst's definition appears to equate student performance with program performance. Some disadvantages of this view are described by the PDK Committee: (a) The evaluator's task is limited to breaking down broadly stated objectives into operational sub-objectives which may or may not reflect the original intent of the program. (b) The evaluator's attention is focused on student behavior even though a measurement of some other function, such as a new staffing procedure, might be what is needed. (c) Instead of using student performance as a feedback mechanism for the purpose of formulating recommendations for improvement, practitioners are forced to view evaluation as a terminal event for rendering judgments of program success. More recently, Stufflebeam and Shinkfield (1985) have listed other effects caused by this conceptualization: (a) It tends to stultify the progressive evolution of the program by refusing to revise or expand the chosen body of goals and objectives. (b) It restricts objectives used for evaluative purposes to those that describe observable or easily quantifiable criteria, and in doing so, ignores important program effects that may be assessed in a different manner. (c) This method cannot measure the degree of implementation of program activities and innovations. (d) It ignores important issues such as the performance and attitudes of program management and staff. The examination of program objectives is an essential component of evaluation, but data relative to students' attainment of objectives is incomplete: (a) It does not provide enough information to effectively judge the quality of an entire program. (b) It does not necessarily produce information that will assist in improving program performance. (c) Such objectives represent only a small portion of program objectives which must address staffing, management, budgets, and other logistics. The problem in focusing solely on the progress of program recipients is that the scope of investigations is too narrow to accommodate the information needs of program sponsors and managers.

The third early description of program evaluation is based on professional judgment. Several methodologies have been established from this point of view: (a) Visitation procedures by accrediting associations employ a team of professionals who visit program sites, observe, and judge the program. (b) A team of specialists from different disciplines collect survey data and produce evaluative conclusions. (c) Panels are organized by funding agencies to review and judge program proposals using different categories such as significance, design, personnel, and budgetary efficiency. (d) Examinations are required of doctoral candidates who must state their conclusions about a particular issue(s) and defend it orally. The PDK Committee lists several disadvantages of the professional judgment approach: (a) questionable reliability and objectivity of those assigned to judge the program; (b) inability to apply ordinary scientific prudential measures; (c) inability to generalize findings to program components; and (d) the ambiguous quality of the data and criteria used for evaluative judgments. Although professional judgment can be a valuable evaluation component when combined with other methods, it is not robust enough to independently identify problems and solutions or measure the overall success of an entire program.

Conclusions about Early Descriptions of Program Evaluation

Early descriptions of program evaluation tend to focus on activities that are important to only part of the evaluation process. The PDK Committee proposed a more holistic approach to educational program evaluation. They define evaluation as ". . . the process of delineating, obtaining, and providing useful information for judging decision alternatives" (p. xxv). It is based on four assumptions: (a) The educational establishment is in a constant state of change or flux due to the evolution of objectives within the educational system. (b) The social, cultural, and technological pressures which determine these objectives are interpreted or delineated in part by evaluation. (c) Evaluation devised responses or strategies must be applied to these pressures. (d) Program evaluation is intended to deliver useful information to educational decision makers. Judging from the more recent literature, this definition is appropriate for describing most training programs. Training objectives are often modified as a function of evaluative data, and the sole mission of training program evaluators is to deliver information to decision makers for the purpose of improving training effectiveness.

Literature on evaluation conducted outside of education also helped to formulate new and broader views of educational program evaluation. Suchman's (1967) work was cited by the PDK Committee as being especially useful for servicing action programs in the areas of social welfare. Quade (1967) indicates that program evaluation should include operations research, systems analysis, and cost/benefit analysis. His description is directed to military decision making. These ideas contributed to a more wide ranging and realistic view of program evaluation.

Early descriptions of program evaluation tend to view the field as a process that involves judging the value or quality of a single product. Scriven (1981) uses Consumer Reports as a reference to examples of product evaluation. Viewing programs as products or individual entities ignores the complexity and interrelationships among program components. In summary, there were some major problems with early evaluation techniques. For example, two general problems appeared to result from a misconception or an inadequate description of program evaluation; i.e., (a) Evaluators had too narrow a focus that included only one or two components of program evaluation, and (b) they attempted to develop an evaluation philosophy by borrowing principles from related disciplines instead of redesigning them to suit the needs of program evaluation.

Advancements in the Concept of Program Evaluation

Since the publication of the PDK Committee's report, the concept of program evaluation has expanded to include a broader scope of activities and responsibility. Its purpose has evolved from one that focused on a simple assessment of a program's overall success to a forward-looking, broad-ranging scope of inquiry aimed at an ongoing improvement process that continues throughout the life of the program. For example, four general requirements for program evaluation were published by the Joint Committee on Standards for Educational Evaluation in 1981:

(a) Evaluations should communicate the strengths and weaknesses of a program to those who take part in its implementation and operation processes, and provide them with recommendations for improvement when possible.

(b) Plans for evaluation should be feasible whereby procedures for data collection and analysis will draw on available resources and will be minimally disruptive to program functions.

(c) Evaluations should be founded on explicit agreements to ensure that the rights of all concerned parties will be protected and that findings will not be compromised.

(d) Evaluations should clearly describe the program as it evolved in its context and reveal strengths and weaknesses of the evaluation design, procedures, and conclusions. This schematic framework demonstrates the magnitude and complexity of modern program evaluation.

The ERS has defined several general categories of evaluation and program-planning activities, some or all of which will be applied to a particular evaluation effort depending on the purpose and intentions of those who commission and execute the evaluation (ERS Standards Committee, 1982). The activities include front-end analysis; evaluability assessment; formative evaluation; summative evaluation; and meta-evaluation (evaluation audit). A summary of each of these activities is provided below and is supplemented with some additional input from other publications.

Front-End Analysis

Front-end analysis takes place prior to the implementation of a program to "confirm, ascertain, or estimate needs, adequacy of conception, operational feasibility, sources of financial support, and availability of other necessary kinds of support." A substantial part of front-end analysis in planning training programs consists of task analyses for defining the types of expertise (i.e., learning outcomes) that students are to possess after completing a training program. Gagné and Briggs (1979) recommend task analysis methods for analyzing human performance and training needs prior to program development. Reigeluth and Merrill (1984) devised a procedure for conducting task analysis that considers both the set of skills to be acquired and the higher-order cognitive aspects of the skills. Similar procedures are used by instructional designers to develop specifications for training.

Rossi and McLaughlin (1979) emphasize the need to work with program administrators to establish clearly defined and agreed upon evaluation objectives prior to an evaluation. Evaluation objectives describe what the evaluation is to accomplish and should not be confused with program objectives which describe what the program is to accomplish. Front-end analysis should be employed as a first step in program planning to flesh out (a) what program

recipients (students) are to gain from participating in the program and (b) what resources and facilities will be required for program development, implementation, operation, and maintenance.

Evaluability Assessment

Evaluability assessment is used to determine the degree to which a program can be effectively evaluated. It may encompass inquiries into the technical and political feasibility, determine how evaluation conclusions would affect funding or program changes, and assess the degree to which the program has been implemented as planned.

The immense cost that accrues in the effort to evaluate large programs and the historically high rate of failure and nonutilization of evaluation findings have made evaluability assessment an important component of program evaluation. Stufflebeam (cited in Brandt, 1978) has stated that one reoccurring problem is that "we [evaluators] gather too much information or the wrong information" during program evaluations. Assessing the degree of program evaluability enables the evaluator to identify the questions that can be answered at a reasonable cost and differentiate them from those that cannot be answered or can only be answered at an unreasonable cost (Nay & Kay, 1982).

Evaluability assessment was developed by Wholey and the Program Evaluation Group at the Urban Institute in the early 1970s (Rutman, 1980). The procedure was designed as a result of the discovery of serious discrepancies between the recommendations of policy analysts and those of evaluators. Analysts' rhetorical descriptions of program goals, functions, and outcomes did not accurately represent data collected in the field which described the actual working environment of the program. Rutman has applied evaluability assessment to several federal and provincial programs in Canada including Job Creation, Manpower, unemployment insurance, social assistance parole, and environmental protection regulations. He claims that evaluability assessment can improve the relevance of evaluation by pointing out those aspects of the program which meet the preconditions of evaluability and those that may not benefit from evaluation. The intent is to "identify particular program components and specific goal/effects that meet the preconditions of evaluability" (p. 88). Strosberg and Wholey (1983) have identified eight questions that an evaluability assessment attempts to answer:

1. What resources, activities, objectives, and causal assumptions make up the program? . . . .

2. Do those above the program managers at the departmental level, and in the Office of Management and Budget, Congress, and the General Accounting Office agree with the program manager's description of the program? . . . .

3. To what extent does the program have agreed-upon measures and data sources? . . . .

4. Does the description of the program correspond to what is actually found in the field? . . . .

5. Are program activities and resources likely to achieve objectives? . . . .

6. Does the program have well-defined uses for information on progress toward its measurable objectives? . . . .

7. What portion of the program is ready for evaluation of progress toward agreed-upon objectives? . . . .

8. What evaluation and management options should management consider . . . (p. 67-68)?

Note that questions 1, 2, 5, and 6 can be partially answered by a thorough front-end analysis.

Host (cited in Strosberg & Wholey, 1983) describes three key conditions that a program should meet which can be verified using evaluability assessment:

> Condition 1: Program objectives are well defined, i.e., those in charge of the program have agreed on a set of realistic, measurable objectives and program performance indicators in terms which the program is to be held accountable and managed. Condition 2: Program objectives are plausible, i.e., there is evidence that program activities are likely to achieve measurable progress toward program objectives. Condition 3: Intended use of information is well-defined, i.e., those in charge of the program have agreed on how program performance information will be used to achieve improved program performance (p. 66).

The three conditions described above can be partially satisfied by conducting front-end analyses. Referring to Condition 1, objectives and performance indicators should be established at the program's planning stage. Condition 2 may be met by examining and comparing the activities of similar programs that have been successful. Finally, Condition 3 points out how important it is for evaluators to form close ties and agreements with program decision makers to facilitate the dissemination of evaluative information in a way that will promote formative improvements as the program matures.

Many problems in program evaluation can be attributed to an insufficient description of programs and their expected outcomes during implementation (Siegel & Tuckel, 1985; Strosberg & Wholey, 1983). Other problems arise due to poor communication or working relationships among program evaluators and key decision makers (Brandt, 1978). Evaluability assessment can be viewed

as an extra effort to fully describe the program and assess the feasibility of collecting and utilizing evaluative information. An evaluability assessment of an existing program can provide a strong foundation for the implementation of a full-scale evaluation system. However, it may not be useful as an overall cure for incomplete program planning or inadequate front-end analysis. Once a program's funding, resources, and operating components have been established, they may be difficult or even impossible to change for the purpose of increasing evaluability.

Wholey (1987) describes four problems that may be detected by evaluability assessment before an evaluation is attempted: (1) lack of definitions of problems to be addressed by the program, program interventions, and expected outcomes; (2) lack of clear logic for testable assumptions linking expenditures of resources to stated outcomes; (3) lack of agreement on evaluation priorities and intended uses of evaluation findings; and (4) inability or unwillingness to act on evaluation findings. Wholey (1987) cites an example of the success of evaluability assessment when it was applied to the Tennessee Department of Public Health. He describes three benefits to this program: (1) an agreement of evaluators, key policy makers, managers, and staff on the theory underlying the program (intended inputs, activities, outcomes, and assumed course linkages among those components); (2) agreement on objectives and their indicators of attainment; and (3) the decision to add an interim report to the evaluation for use by the state's budget process. This account provides evidence that evaluability assessment can be successfully applied to extant programs, but it does not describe the costs associated with the assessment and the program improvements that followed. It is possible that efficient planning for evaluation prior to implementation of the program would have precluded the need for post hoc evaluability assessment. Conscientious planning for evaluation is likely to save a great deal in terms of (1) possible disruption to the operation of the program and (2) the associated costs of adjusting the program to increase its evaluability.

It is interesting to note that most proponents of evaluability assessment recommend its application to programs that are already in operation or those that are in the process of being implemented. However, program evaluators are seldom able to shape programs to maximize evaluability once they are operating (Heilman, 1980).

Cohen, Hall, and Cohodes (1985) have proposed an a priori method they call "evaluation readiness" that seeks to prepare new programs to yield useful evaluation findings and to prepare evaluation staff to respond in a timely fashion with information helpful to program managers and planners. They claim that evaluability assessment recognizes the importance of "essential ingredients for successful program evaluation, but it places relatively greater emphasis on documenting their absence rather than assuring their presence" (p. 316). Evaluation readiness requires a rigorous approach to evaluation that involves the establishment of a "formal program definition" and a description of data inventory prior to full implementation. The program definition requires three steps. The first is to produce an overview that describes the logical flow of program operation in terms of program processes, outputs, and outcomes. Next,

program information must be gathered that answers three questions: (a) Are there multiple management or organizational perspectives that require separate delineation or will a combined-perspectives approach be most useful? (b) Are program objectives linked together, and do they reflect the operational flow, or do they need revising? (3) Are originally stated objectives measurable, or do they require revision? Finally, program objectives must be organized, and their causal links to each other must be defined in terms of process, output, and outcome.

A "data inventory framework" is proposed by Cohen, Hall, and Cohodes (1985) for determining the evaluation readiness of programs. It has four steps:

> Step 1: Constructing Program Performance Measures . . . .
> Step 2: Identifying Data Requirements [each measure] . . . .
> Step 3: Assembling the Data Inventory Framework [a table format that describes the operational relationship among objectives] . . . . and
> Step 4: Reviewing Results with Program Management to assure the validity of the framework and to enhance its utility for decision making . . . (p. 319-321).

The tasks specified for evaluation readiness are similar to those described for evaluability assessment, but its emphasis on an a priori approach is commendable in view of the expense involved in attempting to force extant programs into a format that meets the requirements for evaluability. Because evaluation is an essential component of programs, it is reasonable to plan for program development, evaluation, and program improvement prior to their implementation.

Formative Evaluation

Formative evaluation includes analyzing and testing the processes of a program in order to make modifications and improvements (Joint Committee on Standards for Educational Evaluation, 1981). Activities may include an analysis of management strategies and interactions among personnel including personnel appraisal, attitude surveys, and other observations. This process review requires the evaluator to work in close association with decision makers during the implementation and operation of the program. There is probably more information available on formative evaluation than any other form of evaluation.

Scriven (1981) states that formative evaluation takes place during the development of a program and is conducted for the inhouse staff. The general goal of formative evaluation is to define the program's environment and describe its processes during implementation prior to becoming fully operational. An adaptive change should take place during implementation that will move the program toward a better fit with its context and expectations of its staff and other key decision makers (Nielsen & Turner, 1983). This is an evolutionary process whereby the evaluation methodology is modified as the ongoing program matures.

There is some disagreement about the duration of formative evaluation. Scriven (1986) states that a formative evaluation report must be prepared while there is still time and resources for improvement, and that it should assist in the development process. This suggests that formative measures should terminate after the program becomes fully operational. However, it seems counter intuitive to completely abandon formative measures and attempts to improve programs after they have become operational. Ceasing all efforts to formatively evaluate programs would be acceptable only if one of two very unlikely conditions were met: (a) The program reaches a point of perfection whereby no further improvement is possible. (b) After the program becomes fully operational, it can no longer be changed to facilitate improvement. Nielsen and Turner (1983), Cronbach et al. (1980), McClintock (1984), and others maintain that formative evaluation must continue throughout the life of a program if it is to operate at maximum efficiency. Given the dynamic quality of most program environments, providing for continuous formative measures that facilitate improvement in program efficiency is sound advice. For example, changes in training program components (e.g., switching from classroom instruction to computer-based instruction) should be tracked to determine their effect on student performance and program completion rates. If changes within programs are not tracked across program cycles, there will be no way to determine their overall effect on the program and no way to assess their advantages and impact on training outcomes.

Chelimsky (1985) suggests that the goals of evaluation are concerned with "relating program activities to program effects in a way that will be useful for a broad array of information needs" (p. 489) and that causes of problems must be detected prior to prescribing solutions. However, in Scriven's (1981) definition of formative evaluation, he maintains that "analytical evaluation . . . may or may not involve/require/produce causal analysis, so the connection between evaluation and causation is pretty remote . . ." (p. 63). This position may be misleading because it is unreasonable to believe that prescriptions for program improvement be based on correlational data or qualitative judgment alone. Also, considering the large volume of evaluation literature that deals directly with inferential analyses and hypothesis testing, this interpretation appears not to be well supported. Finally, it is fairly obvious that program decision makers need to make some assumptions about causes if they are to formulate solutions to problems. The purpose of exemplifying apparent contradictions in the literature is to point out the danger of developing evaluation strategies based on the interpretation of a single author's opinion.

Evaluation researchers have expressed the need for close observation of programs during their implementation. Analyses of the implementation process serve to describe differences between the intervention activities that are carried out and those that were originally planned (Nay & Kay, 1982; Scheirer & Rezmovic, 1983; Leithwood & Montgomery, 1980). Early analyses facilitate program decision makers' efforts to formatively shape the program as it matures. Scheirer and Rezmovic (1983) recommend that efforts be made to determine the degree of program implementation prior to drawing conclusions about the effectiveness of its components. They describe implementation as "the extent

of change that has occurred at some particular time toward full, appropriate use of the target innovation" (p. 601). The procedures they describe for assessing the degree of implementation are derived from 74 case studies. The most common methods used were interviews and questionnaires that targeted program staff and management personnel. Program staff are favored over management as the target population for interviews and questionnaires. This view is supported by Nay and Kay (1982) who emphasize that staff members located at direct-intervention levels usually produce the most concrete descriptions of program activities.

Nay and Kay (1982) describe four steps for formative development of measurement procedures: (a) defining and selecting characteristics (i.e., performance indicators) to be measured; (b) defining and selecting a metric and scale; (c) selecting or creating an instrument; and (d) estimating the degree of accuracy of data to be obtained. Boruch, Cordray, Pion, & Leviton (1983) have suggested the use of pilot evaluations to resolve disagreements concerning which questions should be addressed, which methods to use, and the quality or usefulness of the information to be collected. Pilot evaluations may be useful in describing formative evaluation tools such as the language to be used in future instruments and evaluation reports.

Worthen and Sanders (1987) have advocated a unique method of evaluation referred to as "adversary-oriented evaluation." This technique may elicit similar levels of motivation and creativity (in evaluators) as competitive sports. Two evaluation teams are randomly assigned positions as program advocates or adversaries. The teams write and exchange reports and prepare written rebuttals. The rationale for this technique is to illuminate both positive and negative aspects of the program, broaden the range of information collected, and boost the interest of intended audiences - "everyone loves a contest."

In adversary-oriented evaluation, opposing views are incorporated into the evaluation goals design so that pros and cons are argued openly. The strategy of anticipating criticism and preparing a strong defense is similar to that which is employed in politics and law. Worthen and Sanders (1987) state that the adversary technique is useful when certain conditions are met: (a) The object of the evaluation affects many people (e.g., large government programs). (b) Controversy about the program has created wide interest. (c) Evaluations are external (i.e., conducted by people outside the program). (d) Clear issues are involved. (e) Resources are available for additional expenses required by adversary strategies. The last point appears to be the "catch" in this technique. The allocations of program funds for evaluative activities are seldom overgenerous. A program that incorporates two evaluation teams will probably cost about twice as much as one that employs a single team. However, a pilot evaluation using adversary teams may produce findings that would be of great importance to the design of long-term evaluation procedures and, therefore, outweigh the extra costs. Employing small groups of graduate student interns to conduct an adversary-oriented pilot evaluation could be cost effective. Such an activity would likely uncover useful evaluation issues and provide students with valuable field experience.

Summative Evaluation

Summative evaluation corresponds to the goal of describing how well an entire program has met its planners' and sponsors' expectations. It is intended to produce information for major decisions about program continuation, expansion, or reduction. In summative evaluation, program performance indicators must be chosen that reflect the overall impact of the program on those who it is intended to serve (Scriven, 1967; Stake, 1967).

Scriven (1986) strongly recommends summative evaluation of a program after it has become operational to determine if it is meeting the goals it was designed to meet. He stresses that summative evaluation of a program is conducted after completion of a particular program cycle or the entire program, and also that it is for the benefit of some external audience or decision maker (e.g., funding agency, historian, or future possible users). Worthen and Sanders (1987) warn that summative evaluations should not be done on new programs that have not had time to be debugged. They advocate that evaluation goals be formative in nature until program processes have become sufficiently mature and stable. Their point is well made--how can one determine the attainment of program goals until one has determined the degree to which the program has been implemented and is fully operational? A program that has not reached a fully operational stage is not ready for summative evaluation.

Auditing shares some characteristics with summative evaluation since both activities involve a monitoring task that compares what has been proposed with real world outcomes. Chelimsky (1985) identifies some of the characteristics shared by auditing and summative evaluation. She indicates that the purpose of an audit is to verify the correspondence between the matter under investigation (e.g., a resource expenditure) and some standard of operation of performance. The emphasis of an audit is on accountability, and the measure of performance is based on normative information described by a particular standard. Summative evaluation is similar because it attempts to determine if the program has attained the terminal objectives it was planned to attain. This approach requires that the program's performance be judged by comparing it to (a) performance goals developed during the planning stage or (b) the performance of similar programs.

Chelimsky (1985) describes several tasks and methods used by both auditors and evaluators: (a) defining success based on the correspondence between program outcomes and objectives, (b) identifying deficits, (c) using criterion-referenced analysis methods, and (d) recording data in a chronological fashion. These procedures are important to those preparing a summative report but are of less value to the formative tasks of maintaining and improving a program. Those planning a summative evaluation or who are primarily interested in program accountability may benefit from the knowledge and methods of auditors.

Sherrill (1984) compares "outcome" (summative) evaluation with cost-benefit analysis and encourages evaluators to combine the two methods to avoid the shortcomings that are inherent to each when they are employed independently. He maintains that outcome evaluation provides for estimations of future outcomes

but yields nothing about the values and costs of the outcomes. Cost-benefit analysis provides information on future monetary benefits and costs of outcomes but is not as effective for analyzing the outcomes themselves. Sherrill claims that using both methods together can yield results that are essential to program decision making.

There remain some strong differences of opinion concerning summative and formative evaluation among even the most noteworthy professionals. For example, Cronbach et al. (1980) asserts that evaluations are intended to serve a political function and should provide formative information that supports negotiation rather than making summative decisions. They stress that there are many dangers in drawing summative conclusions about program efficacy. This view is strongly criticized by Scriven (1986) who calls for a greater emphasis on accountability. Scriven maintains that program improvement requires summative measures, and that decisions must be made whether to admire or condemn whole packages of attributes ". . . whether the packages are projects, products, or people." Both views are reasonable, but their applicability is dependent on the characteristics and maturity of the target program. Britan (1978) points out that a summative evaluation assumes that ". . . explicit program goals can be isolated . . ." An additional requirement would be that indicators of goal attainment be clear, agreed upon by decision makers, and available to evaluators. Such conditions may or may not be met depending on the maturity of the program, the political climate, and a number of other factors. For this reason, arguments for an emphasis on formative evaluation over summative or vice versa may be useful only in their power to illuminate the problems and responsibilities of evaluators.

Meta-Evaluation

Meta-evaluation is a term that has been used to refer to two completely different endeavors. The first refers to an audit or analysis of the quality of a particular evaluation which will be referred to as evaluation audit. The second refers to a methodology for combining results from multiple evaluations to formulate overall conclusions about a program. The two types of activities are described separately below.

An audit of an evaluation is usually conducted to fulfill the requirements of agencies in coordination or oversight roles. The type of activities involved in evaluation audits performed by the GAO on Federal program evaluations are summative in nature because they examine the quality and validity of evaluation results and conclusions. Evaluation audits in government programs may become increasingly important because many of the evaluations are being carried out by the very firms that produced the systems being evaluated (Cook & Leviton, 1983). House (1986) explains how the need for evaluation auditing is produced both by the lack of documentation critiques and by the political environments surrounding evaluators. Evaluation documents need to be subjected to professional scrutiny by nonbiased experts to help prevent evaluators from succumbing to the temptation to satisfy sponsors with favorable results. Disagreements about validity of evaluation results and possible bias are not

uncommon, especially when they may determine the continuation or discontinuation of a large program that provides jobs for hundreds of people. Stufflebeam (cited in Reineke & Welch, 1986) emphasizes the importance of evaluation audits by stating that ". . . . evaluators are being increasingly required to evaluate their own work . . . . and have come under pressure to ensure the quality of their work" (p. 17). A critique of every major evaluation funded by the Department of Education has been recommended to Congress by Boruch et al. (1983).

A formative approach to evaluation auditing is described by House (1986). He and other evaluators conducted an evaluation audit on the Promotional Gates Program that was implemented in the New York City School District in 1981. The chief task of the evaluation auditors was to assist and advise the program evaluation staff in matters related to evaluation design, techniques employed, and wording of reports on evaluation findings. The principal purpose of this evaluation audit was that the mayor's office, which provided funds for the program, did not entirely trust the objectivity of the evaluators. House advocates the use of evaluation audits to provide "quality control measures" to evaluation projects. In recognition that evaluations must be adjusted to conform to a budget, timeline, or other factors, Reineke and Welch (1986) propose a "client-centered" approach to the evaluation which helps evaluators improve their practice and helps clients make better use of evaluation information. Cook and Gruder (1978) reviewed four projects that applied evaluation audit techniques to summative evaluations. They describe a formative approach which includes an audit of the evaluation procedures during an evaluation of a program and a summative approach which involves an analysis of the evaluation procedures and conclusions after the evaluation has been completed.

The Congressional Budget Act of 1974 directed the GAO to analyze program evaluations prepared by federal agencies. The GAO's more recent criticism of program evaluation as being wholly inadequate within the Federal Government (U.S. General Accounting Office, 1988) seems to indicate one or more of four possibilities: (a) Evaluation audits have not been effective in improving evaluations within the Federal Government. (b) Evaluation audits have not been properly conducted. (c) Results and recommendations have not been followed by program decision makers. (d) Not enough evaluation audits have been conducted to accurately assess the state of program evaluation efforts.

An audit should examine the validity, credibility, utility, robustness, and cost-effectiveness of the evaluation (Scriven, 1981). Scriven emphasizes the importance of estimating differential costs of Type I errors (i.e., detecting effects of program interventions when no effect actually exists) and Type II errors (i.e., failing to detect actual effects of program interventions). The great cost (in both monetary and human resources) of conducting evaluations of large programs makes such estimates important. For example, a high probability of a Type II error is extremely critical if the main purpose of the evaluation is to prove the program is valid and should receive additional funding. The additional expense of conducting an evaluation could exhaust badly needed funds but fail to verify the program's value. A poor evaluation or an evaluation conducted on a

program that is low in evaluability could result in the expending of funds on an unsuccessful attempt to raise new funds! In judging the inclusiveness and general quality of evaluations, some tools have been developed in the form of evaluation criteria such as described by Stufflebeam (1975). Applying such tools to assess the quality of evaluation designs prior to an evaluation may lead to more effective evaluations which would reduce the need for expending further resources on evaluation audits.

Meta-evaluation is performed for a different purpose than the evaluation audit. Meta-evaluation refers to the application of a statistical procedure called meta-analysis (Hedges, 1983, 1986) to program evaluation. By combining and synthesizing the results of multiple evaluations, meta-evaluation provides a synthesis of multiple data sets and makes it possible to formulate more reliable conclusions about program performance. However, meta-analysis procedures are somewhat limited in that they require the results of many evaluations that were conducted on the same program or very similar programs with relatively the same focus (i.e., the same or similar performance indicators). Also, the previous evaluations must have been based on quantitative data where assumptions required of inferential analysis could be reasonably met. Hedges (1983, 1986) has done substantial work in explaining the methodology of meta-analysis and how it can be applied to program evaluation.

Conclusions on the Conceptualization of Program Evaluation

There is little doubt that program evaluation includes a much broader scope of inquiry than was assumed by early measurement-oriented, objectives-oriented, and professional judgment evaluation models. This broad view is immediately apparent in the definition of program evaluation proposed by the PDK Committee and in the evaluation and program-planning activities defined by the ERS. Nielsen and Turner (1983) indicate that the expansion of program evaluation has led to a shift away from the hypothetico-deductive paradigm (i.e., experimental approach) toward a new paradigm that emphasizes the use of multiple methods and procedures for matching evaluation designs to specific program needs and environments. Cronbach et al. (1980) indicates that evaluation is now a more eclectic field that covers a broad range of practices from traditional experimental research to theories of business management and policy analysis. Many of the methodologies for evaluating decision-making processes, policies, and political environments are basically the same as those used to evaluate training and education programs. Although there are some differences, these differences may be overemphasized, and their similarities are probably of greater importance (Nagel, 1985). Program evaluation should be viewed as an iterative process that begins before the program is implemented and continues to provide formative support throughout the life of the program. Evaluators have realized that the inherent multiplicity of programs requires different evaluation activities to be employed based on the program's stage of development and on the needs and expectations of the program decision makers.

The Problem of Decision Making

The results or conclusions of any evaluation are formulated for the purpose of making some type of decision(s). Judging from all the literature examined in this review, the most important function of program evaluation is to provide program decision makers (i.e., staff, managers, and sponsors) with information that will help them take actions that will result in improved program performance. Although the PDK Committee's report outlines a theoretical construct of the decision-making process, it does not describe specific methods that are immediately applicable to program evaluation. This lack of specification is understandable, because according to Thompson (1982), systematic methods for decision making are relatively new and were considered for application to program evaluation only in the last decade. In describing one particular methodology, he mentions that by 1965 a comprehensive mathematical procedure for decision analysis was in existence, but the technical level was too formidable for most evaluators and was, therefore, not put to use.

A decision based on evaluative findings may be as broad and all-encompassing as the decision to continue or discontinue an entire program or as specific as the recommendation to allocate funding for the purchase of a microcomputer. Evaluators often lack a clear and useful conception of the scope of program decisions, and this clouded perspective hinders them in formulating appropriate strategies for generating the information needed to help decision makers improve their programs.

The mechanics of evaluation often require the evaluator to apply highly technical data analysis methods (e.g., multivariate statistical analyses), but establishing functional links between evaluation results and decision makers' needs requires a completely different type of expertise. Skills that are more closely related to those possessed by counselors, attorneys, and corporate managers are more valuable for interacting with decision makers. This task can be very complex because programs usually possess a multiplicity of decision makers, interdependent decisions, and different kinds of criteria. If evaluation results are to be useful, the evaluator must analyze and delineate decision chains and structures, describe how the different forms of evaluative data reflect program performance, and assist decision makers at every step of the decision process.

Program evaluators will benefit from knowledge collected from several sources: (a) recommendations of evaluation practitioners and researchers for facilitating the use of evaluation findings; (b) the theories and methodologies of policy analysis; (c) the analytical and computational formulas referred to as "decision theory" or "decision analysis," and (d) theories and methods applied to personnel management. The PDK Committee provides some useful ideas that address the theoretical implications of decision making. A method of organizing categories of decisions is presented along with descriptions of different types of decision-making environments and factors that affect the decision process. The PDK Committee's views on decision analysis are derived from a book by Braybrooke and Lindblom (1963). Strategies and computational formulas from

decision theory are too extensive to be comprehensively covered in this review. Therefore, only a general discussion of decision making in program evaluation will be presented.

The PDK Committee recommends that program evaluators concentrate on five main tasks of decision making: (a) Identify the decision makers within the program. (b) Define the decision questions they must answer. (c) Identify the decision alternatives that must be considered. (d) Define the criteria to be used in judging alternatives. (e) Describe the projected timing of the steps in the decision process. The four stages of decision making that encompass these tasks are (a) becoming aware that a decision is needed; (b) designing the decision situation; (c) choosing among alternatives; and (d) acting upon chosen alternatives.

The decision setting (i.e., program environment) will determine the type of strategy most appropriate for initiating change in a program. Braybrooke and Lindblom (cited in Stufflebeam, et al., 1971) have identified two variables that describe a decision setting: (a) the level of "information grasp" (i.e., how well the decision makers understand the features or data relative to the problem at hand) and (b) the "degree of change" that the decision maker perceives will occur as a result of the decision. It is the evaluator's job to increase the level of information grasp and provide decision makers with an estimate of the degree of change for each decision.

Some very serious obstacles that block effective decision making are present in large or complex program environments. One of the most difficult to deal with is "nesting" whereby multiple decision makers are involved in the same decision. This situation creates an interdependency among decisions that increases their complexity as a function of the number of decision makers and levels of authority involved in the decision process. The evaluator is responsible for analyzing the population of decision makers within a program. This analysis involves several tasks: (a) Identify the different value positions of decision makers, e.g., a training program manager who is striving for a reduction in dropout rate vs. a program sponsor who is ready to cut off funds unless overall trainee performance improves. (b) Describe the different criteria and different goals that may emerge from different decision makers. (c) Describe alternative actions to achieve the different goals. (d) If possible, estimate success probabilities for each alternative. (e) Delineate an optimal compromise alternative within each conflicting goal.

The interdependency of decisions is sometimes based on a contingent series whereby previous decisions have a bearing on future decisions. For example, a reduction in training hours may not affect students' acquisition of a particular skill, but it may reduce the time they are able to retain the skill. Therefore, a decision to reduce training hours may not affect decisions relative to testing criteria, but it may affect decisions about retraining cycles to maintain skills. Higher-order decisions may also modify lower-order decisions and their alternatives For example, the implementation of computer-based training modifies the range of different types of instruction available. This change, in turn, modifies the

decision alternatives relative to developing new training sequences. Evaluators should delineate the cause-effect relationships within each series of decisions. This structuring can be done by describing a hierarchical decision structure composed of multiple levels of decision categories and delineating the effect of each level on the next level down in the structure. Such a model will facilitate decision makers' and evaluators' understanding of how the program will be changed before any decisions are finalized.

The PDK Committee outlines several conclusions concerning decision making: (a) The relationship between evaluation and decision-making roles is symbiotic. (b) The evaluator's role in servicing decision makers varies according to decision settings. When the information grasp of decision makers is high, and the decisions are perceived as having minimal impact on the program, decision makers need less help from evaluators. But, when the information is difficult to understand, and the decisions are perceived as having a large impact on the program, evaluators must play a more active part in the decision-making process. (c) Program planning decisions need to be made in most settings, and the consequences of these decisions will determine the nature of future decision-making settings. (d) Evaluators should not make or implement decisions, but they should assist decision makers at all stages of the process. (e) Evaluators' involvement in planning decisions is systematic while it is ad hoc with respect to structuring and implementing decisions. (f) Evaluators must possess a broad range of abilities in interpersonal and technical skills to successfully serve the needs of decision makers. (g) To make effective decisions, *evaluation must be a cooperative effort* that incorporates evaluators, administrators, program staff, and a variety of consultants or subject matter specialists in a team effort.

The essential ingredient for decision making is information. The quality of this information can increase or decrease the level of difficulty for making a decision and determine the degree of likelihood that the alternative chosen will prove to be the best choice. There are several qualities that make information useful. Information should: (a) be clear and understandable to decision makers; (b) be timely in terms of evaluation deadlines and program revision plans; (c) carry the same meaning for all decision makers; and (d) be verifiable in terms of accepted concepts and supportable measures.

Whatever approach the evaluator takes in the decision-making process, the first task is to identify the individuals who are motivated, understanding, able, and have the authority to take actions based on evaluative information that will lead to program improvement (Nay & Kay, 1982). The evaluator's next task is to establish a good working relationship with these individuals and other program staff. This rapport development includes "learning the language" of program personnel and working out a common context on which evaluation results can be reported and understood. The benefits of evaluators working closely with program decision makers cannot be overemphasized. Roberts-Grey, Buller, and Sparkman (1987) stress that evaluators often lose focus on program needs by concentrating on running tests and collecting and reporting data instead of assisting decision makers in alleviating problems and improving

program efficiency. Adjustments in program components are largely based on the degree to which stated objectives are being met (Stufflebeam & Shinkfield, 1985); therefore, a common understanding among evaluators, decision makers, and program staff on these objectives, their criteria, and indicators of performance will determine the effectiveness of the decision making process.

Thompson (1982) has devoted an entire book to the application of decision analysis to program evaluation. He describes decision analysis as a determination of options that best serve the interests of decision makers. According to Thompson, decision analysis provides a holistic perspective that can lead to good evaluation decisions. He explains that decision analysis is geared for a prospective approach--looking ahead to consequences of future decisions--rather than a retrospective approach--looking back at how well a program has done in fulfilling its initial goals. This is an important point, because the early program evaluations that failed tended to use the retrospective approach. The primary function of early program evaluations was to measure the extent that a program had accomplished what it was designed to accomplish. The evaluator's role was somewhat passive in relation to decision making and ended at the point of presenting the evaluation results. The methods used in decision analysis can be applied to evaluation activities from the program planning stage throughout its implementation and operation.

One of the most difficult tasks in decision making is the delineation of evaluative results to describe their relation to particular decisions and decision alternatives. Results of evaluative observations and their relation to decision alternatives must be presented in a form that is interpretable by decision makers. Thompson (1982) describes the use of "decision trees" for depicting decision-making structures. Decision trees are diagrams that identify decision points (i.e., locations within a program's development process where decisions need to be made). Probability, cost, and other values are assigned to each decision outcome and these values can be used to "solve the decision tree" which results in describing optimal choices. The values assigned to each outcome are derived from various sources including evaluative measures, cost records, and program goals and objectives. Decision trees provide (a) a systematic method for deriving the decision-making structure; (b) a method for communicating the relationship between evaluative data and program goals; and (c) a tool for decision makers that describes the relationships among multiple decisions and an estimated impact of choosing decision alternatives. Decision trees may be effective if precise measures of probability and costs can be obtained. However, the complexity of decision trees and the computations involved in their structure assume that values are accurate. Even a small degree of error in numerical values may render a decision tree useless, or even more serious, may lead to erroneous conclusions.

As an alternative to constructing decision trees, some evaluation researchers have proposed rule-based decision models. Ross (1980) has advocated a procedure for generating "decision rules." This approach is similar to the one proposed by Roberts-Grey, Buller, and Sparkman (1987) for developing "evaluation data rules." Both methods employ rules to summarize information about the

value of decision alternatives. One of the major problems in program evaluation involves the task of sorting out useful elements from a huge volume of information. Decision rules help to restrict data collection to appropriate elements and provide a mechanism for condensing and organizing information into a set of manageable chunks. The broadest possible configuration of a decision rule would consist of three alternatives--terminate the program, continue it, or modify it. Of course, program decision makers require a host of rules that are much more specific and target particular aspects of the program such as personnel training, management processes, and program interventions. Because decision makers will be most knowledgeable of their own domain, Ross (1980) recommends that evaluators seek their help in setting up decision rules. An additional benefit of involving decision makers in setting up decision rules is that they will be more inclined to abide by rules that they have played a major role in designing. Such participative management initiatives will reduce the variance among evaluators' and decision makers' interpretations of program needs.

Some potential problems should be mentioned in the application of decision rules. Ross (1980) states that setting up decision rules too early in the program may invoke closure on the development of decision alternatives thus limiting creative solutions to problems. He offers a systematic solution to this problem. First, develop an "initial primitive set" of decision rules using hypothetical data during the program planning stage. Second, develop a more detailed set using operational definitions keyed to the data collection instruments as they are developed. Third, decide on a final set of rules during the data analysis. This strategy will ensure that the decision rules evolve along with the program as it matures. The fact that programs do change significantly over time is important in that any processes that are integrated within the program, including evaluation, must be dynamic in nature if they are to remain effective (Nielsen & Turner, 1983). If evaluators wait too long in setting up decision rules, they will likely get trapped into formulating rules and attempting to identify alternatives at the same time that they are laboring to communicate findings to decision makers.

The method used to document and disseminate evaluative information to decision makers has a significant influence on information utilization. Pollard, Cooper, and Griffin (1985) define evaluative information as "written material in human- or machine-readable form, that pertains to plans, activities, and results of the project" (p. 161). They point out that the very process of writing down plans and decisions is important because it "encourages, if not forces a thorough consideration and articulation" of the information. Haskins (1981) has proposed a special notational system for clarifying documentation of program interventions and evaluation activities. Such conventions for documentation may help to establish a common language among evaluators and decision makers provided the notation can comprehensively describe the evaluation design, evaluation data, and decision structure. A standard notation can also reduce the disruptive effect of staff turnover and make the task of transferring information from interim reports to final reports easier.

Pollard, Cooper, and Griffin (1985) describe six types of documentation that cover different aspects of a program and are important for evaluation:

(a) Planning documents include objectives and tasks described for each phase of evaluation.

(b) Design documents describe the relation of evaluation objectives to program components, variables of interest for evaluation, measuring instruments, and populations to be studied.

(c) Data collection procedures describe how data will be obtained, the selection and training of data staff, and the scheduling and monitoring of data collection tasks.

(d) Data processing and analysis includes the processing steps to be used for data cleaning, validation, file construction, and statistical analysis.

(e) Integration and summary defines and outlines program events, discussions of problems/solutions encountered, and provides a comparison of what was planned with what was done in the actual program including time, implementation, staffing, resources, and budget.

(f) General project documents are accounting records, cost reports, personnel files, descriptions of contracts and agreements, the program proposal, and an index of available documents.

The collecting, storing, dissemination, and documentation of program data can be made easier, faster, and more efficient by using computerized systems. Computers are especially effective for ongoing formative evaluations. Once the type and sources of data have been determined, computers can be used for systematically analyzing information collected from each program cycle. A well-designed data base can also facilitate the combining of many data sets for conducting meta-evaluations. Local area networks that link microcomputers so data can be transferred back and forth are especially effective for keeping sponsors and staff members up to date on large programs. Many software packages that have been designed for these applications are inexpensive and require minimal personnel training. Most program agencies have the facilities for such systems but fail to consider using them for automating evaluation tasks. The savings through the use of computerized evaluation techniques should be investigated for any long-term program evaluation plan.

Roos, Nicol, Johnson, and Roos (1979) suggest using extant data bases for evaluative purposes. They describe a case study where an administrative data bank was employed for an evaluation of the Manitoba Health Services Commission. They conclude their report with a list of advantages for exploiting such data banks: "(a) wide coverage to facilitate generalizability; (b) a large enough N to permit a number of simultaneous controls in data analysis; (c) a long time-series of data to allow analysis before and after an intervention; and (d) the potential for combining files to facilitate the analysis of intervention, the construction of comparison groups, and the generation of histories for individual respondents" (p. 252).

Program evaluations often generate tremendous volumes of data that must be reduced to a more manageable format before being reported to decision makers. Processing large volumes of data manually leads to increased human error which ultimately reduces accuracy and reliability of analysis. Nagel (1986) advocates the use of microcomputers for evaluation tasks. He describes available software designed as an aid to decision making, establishing relations among evaluative factors, gathering information, and even teaching evaluation skills to prospective practitioners. Because of their speed and accuracy in handling large volumes of complex data, computers are strongly recommended for the practice of program evaluation.

Providing program decision makers with useful and understandable evaluative data for the purpose of improving program performance is the most important task an evaluator is likely to perform. A large range of decisions, decision alternatives, and decision outcomes based on the evaluation plan should be organized into a meaningful structure to facilitate the decision-making process. The decision structure and the rules and guidelines for making decisions based on evaluative data should evolve and mature as the program matures. Because the key to effective decision making is clear, relevant information, evaluators should avail themselves of the most advanced and effective information-handling technology. This technology includes skills associated with developing close, working relationships with decision makers, decision analysis techniques, and state-of-the-art data management systems (e.g., hardware and software). Finally, clear and comprehensive documentation of program development and evaluation is necessary to effectively collect, analyze, and manage data for recommending decisions that will improve programs.

The Problem of Values and Criteria

Determining the type of data needed to assess program performance and identifying criteria that are valid and reliable are formidable problems for evaluators. Most objectives-oriented methods for evaluating programs do not question the validity of objectives themselves. Instead the focus is on determining the degree of congruence between program objectives and outcomes. No adequate methodology appears to exist for combining multiple values or synthesizing new values, even though such a determination constitutes one of the most important tasks the evaluator performs. This problem arises from the pluralistic nature of programs where multiple values of decision makers may be present in the same program environment. Data interpreted by different value standards may give rise to antithical evaluative conclusions.

It is apparent that values must be clearly described prior to formulating program goals and objectives. An agreement on what constitutes progress toward, and attainment of goals and objectives must be established early in the program. Every adult has established a set of core values which is constantly used when making judgments and forming attitudes. These core values differ among individuals, and they rarely undergo substantial changes in

short periods of time (Petty & Cacioppo, 1981). This is one reason why people's attitudes are not easy to change. The process of a program's formative development during its implementation should result in modifications of goals and objectives (Leithwood & Montgomery, 1980; Nielsen & Turner, 1983; Scheirer & Rezmovic, 1983) but should not attempt to change decision makers' values. Formative development should establish and maintain an equivalent understanding of goals and objectives among program planners, administrators, and evaluators. Effective program modifications can be made as a team effort only if the program goals and objectives fit within the framework of each decision maker's conception of the program's purpose and intent.

The PDK Committee indicates that the decision process takes place for each decision maker as a function of two types of models: (a) a "Type I model" which is the decision maker's idea of what should be and (b) a "Type II model" which is an idealized version of the change process that will bring about the state of affairs described by the Type I model. The Type I model is used for comparison with evaluation results to determine if the program is performing as it was intended. Criteria, as described by the program goals and objectives, will be used to determine if a discrepancy exists between the way the program should be performing (Type I model) and the way the program is performing (as described by the evaluation results). Criteria act as yardsticks for measuring the difference between the Type I model and program performance which is described by the results of an evaluation.

If program performance seems incongruent with the Type I representation, decision makers will attempt to effect a change through their Type II model. A highly skilled and experienced decision maker will formulate a Type I model that is accurate and a Type II model that is feasible and practical. There are two major problems that will produce indecision: (a) The decision maker's Type I model is not defined well enough to make a comparison, and (b) the decision maker is not provided with enough relevant evaluation data to compare the Type I model with the program. The first problem may be partially alleviated by employing a team effort for setting and modifying goals and objectives to take advantage of the greater expertise of the more experienced decision makers. The second problem requires a reexamination of available data and its congruence with program goals.

The Problem of Administrative Levels

Evaluations have traditionally been focused on the measurable effects a program has on recipients. This data-intensive approach is called a "microscopic view," whereas an assessment of the entire program (i.e., its purpose, goals, implementation, maturation, and environment) is called the "macroscopic view." The microscopic approach provides data that helps determine the needs of program participants, but may contribute little to the development and improvement of the program as a whole. The PDK Committee describes several problems that result from using microscopic techniques on the complex environments of educational programs:

(a) The emphasis on behavioral objectives is designed to assess student performance rather than program performance.

(b) It is difficult to aggregate program data in a way that summarizes program performance and establishes a baseline performance rate across different locations and recipients and isolates the effects of different program components.

(c) There is a shortage of methods for analyzing data in accordance with the varied purposes of management at different levels of program administration.

(d) There is some confusion between congruency (the degree of agreement between program objectives and program outcomes) and contingency (goals associated with projecting what might be achieved if certain conditions are met).

(e) There is still a poor understanding of the process of collecting and using baseline data for programs.

A macroscopic view enables evaluators to take into account the interactions among the components that make up the entire program environment rather than concentrating on the program's specific affect on recipients. The PDK Committee explains that early estimates of educational program performance were based on the individual student as the unit of measurement and analysis. Consequently, evaluators have continued to concentrate on the program recipient in the attempt to describe the effectiveness of entire programs even though an approach that covers the entire program in a holistic fashion is necessary if useful and accurate conclusions are to be made. The conclusion is that evaluators need to develop multilevel evaluations that will function in the best interests of each level of the program administration. This tiered approach calls for multiple evaluation strategies to be implemented in the form of information systems.

Information systems typically require (a) data input sources; (b) ordering, sequencing, classification, and analysis procedures; and (c) formats for displaying the processed information. The decision makers at each level of the administration must be consulted in order to specify the expected decision situations to be served by the system if the interface between evaluator and decision maker is to work effectively. The various types of information and the format of evaluation reports must be tailored for each level of administration to be served by the evaluation data base. The term, data base, may refer to a system that employs personnel, computers, and other vehicles capable of data storage, analysis, or dissemination. The data base must not only be technically efficient in data manipulation and processing activities, but it must also contain the right kind of data and deliver this data in a format that is readily interpretable to those who need it to make decisions.

The micro level limits evaluators' scope of interest to subsamples of student performance data. It does not include a system for monitoring and describing

the whole decision-making environment which covers all program processes. The PDK Committee describes some serious problems with the micro view of evaluation:

> When attempting to meet macro information requirements [information on the entire program or partitioned information relative to a particular component] through the use of micro evaluation techniques, the evaluator encounters four major problems: (a) Micro methodology does not yield data needed by higher decision making levels [higher than the instructor] (b) . . . compiling information about individuals that is to be reported to higher levels . . . [creates problems in] preserving individual rights . . . . (c) . . . success criteria vary across vertical [administrative] levels of a system, and attempts to aggregate data from one level to the next are doomed to failure (d) . . . the use of microscopic oriented techniques may result in undue concern with congruency type studies (p. 120).

Information relevant to evaluation is described by the PDK Committee as the aggregation and arrangement of data elements to reduce uncertainty on the part of the decision maker. Various levels of administration encompass populations of decision makers who require different types of information. For example, a state education department may need information that compares school district performance with standards or goals set for a program on self-paced instruction. A school superintendent may want to compare a new grading technique (implemented within the self-paced program) to the previous grading technique. Teachers may need to know what students' attitudes are toward a new teaching style designed for the program.

In addressing the problems associated with administrative levels, the PDK Committee includes nine topics of interest: (a) a theoretical distinction between the basic concepts of information and data; (b) a systems-oriented conceptualization of evaluation; (c) input process and output variables; (d) probability; (e) theoretical implications of systems that service multiple levels of decision making; (f) systems and programs, additivity; (g) information specificity; (h) the need for a data base in systems evaluation; and (i) a strategy for developing multilevel evaluation systems.

Most of the work completed by the PDK Committee concerning the problem of administrative levels represents only the initial identification of problems. These were entirely new discoveries in evaluation research at the time that the PDK Committee was formed, so they were not able to give any solutions or methods that would be immediately usable by program evaluators. Although more recent information on dealing with this problem is available, it will probably be found in more general literature on policy analysis and business administration.

The Problem of the Research Model

The PDK Committee reports that one of the major weaknesses of evaluations on educational programs conducted in the 1960s was the failure of evaluators

to recognize the essential differences between program evaluation and educational research. The purpose of educational research is to generalize the effects of treatments (e.g., instructional methods) across a wide scope of learning environments, but a program evaluation is conducted to assess the effects of treatments within the context and scope of the unique environment of the target program. Research may be conducted within program environments with the intent to generalize findings to other, similar programs. However, the primary goal of a program evaluation is to assess the efficiency of the target program and to transform evaluative findings into information that will lead to the improvement of the program being evaluated.

Researchers seek to partition the environment to extinguish the effects of extraneous variables. For example, educational and training researchers usually choose experimental subjects from a broad population of learners (e.g., K-12 students or college undergraduates). They try to eliminate the effects of variables not under investigation such as differences in school districts, age, gender, and prior knowledge. However, evaluators must be careful to consider the effects of a program within the context of its operating environment and any unique characteristics of its recipient population. They must study the effects of program components in the presence of all variables that will normally be present during the operation of the program. The variables that are important to an evaluation of a particular program are often the same variables that a researcher would attempt to partition out! For example, an evaluation of a U.S. Air Force pilot training program will concentrate on training effectiveness for a unique population of highly skilled officers. Generalizing the effects of this program across a large population of learners is not usually the goal.

In contrast to common evaluation goals, research efforts may be embedded within a program. For example, a new method for assessing aviation skills may be investigated within a pilot training program with the intent of generalizing the findings to other aviation programs. Even basic research (e.g., analysis of spatial-cognitive skill acquisition) may be conducted within new program environments. However, the findings will not be used in the same way as evaluative information collected for the purpose of improving the program. The main point is that evaluators need to be aware of the purpose of each program component and the data collected from that component. A distinction between evaluation goals and research goals is usually possible and should be established prior to conducting major evaluation activities.

Experimental research procedures are often difficult or impossible to execute within the context of programs for the purpose of evaluation. Experimental treatments are typically produced by manipulating an independent variable(s) (e.g., performance feedback) that is thought to affect learning in some way. When groups of learners receive differential treatments, the experimental procedure is referred to as a between-subjects design. In a controlled, experimental environment, this procedure is usually no problem. However, in the environment of a program that is being implemented or is already fully operational, manipulation of treatments and people becomes very difficult or impossible. First, it is usually considered to be unjust to assign different people

to different types of training, because it would be unfair to allow some students to experience more effective training than others. A second problem is related to the expense associated with the application of experimental procedures. Since most large programs are operating in more than one location that may represent diverse environments, the experiment (manipulation of a program treatment) would have to be duplicated in each location in order to determine if the effects are generalizable to the entire program. This is an expensive and very difficult operation. Finally, the rigorous assumptions on which statistical methods are based, and which are essential to experimental research, can seldom be met within program environments.

The PDK Committee points out that "the purpose of research is to provide new knowledge . . . that is universally valid . . . [but] the purpose of evaluation . . . is to delineate, obtain, and provide information for making . . . decisions . . . [that may be] highly particularistic to a specific decision situation, rather than generalizable to many or all settings" (p. 140). Program evaluation attempts to produce information that is valid and useful within the decision-making context of the target program. If the decision-making context is highly generalizable, the purposes and methodologies of research and evaluation may be equated. This latter case is exemplified in some large-scale training programs that include a research and development (R&D) component which operates simultaneously during training activities. The aim of the R&D component is to explore new training methods, systems, or devices. The findings of this type of research may be generalizable to other training programs.

Some of the failures of past evaluations have been blamed on the attempt to adapt a program to suit the assumptions of experimental design or vice versa. For this reason, modified experimental techniques have been developed that have proven to be useful alternatives to methods traditionally used in laboratory environments. The effects of intermediating variables (e.g., variance in instructors and program sites) in the program must be taken into consideration in the evaluation design and analysis process. Standard methods used in research designs to control the effects of other variables such as randomization may be difficult or even impossible in some settings. One alternative is to use quasi-experimental designs and techniques (Votey, 1981). Methods for conducting field studies, qualitative analysis, and other alternatives to experimental procedures have also been proposed. These are described in detail by several evaluation researchers: Alemi, 1987; Boruch, 1975; Britan, 1978; Dobson & Cook, 1979; Filstead, 1981; Greene & McClintock, 1985; Horn & Heerboth, 1982; Kennedy, 1979; Rossman & Wilson, 1985; and Stufflebeam & Webster, 1981.

Because basic research often deals with variables that are new or poorly understood, post hoc methods are often employed. It is not uncommon for researchers to observe the effect of psychological variables on learning prior to formulating descriptions of how and why the effects occurred. In contrast, evaluators need to establish a clear understanding of what is expected of the program components a priori in order to make accurate estimates of effectiveness during program cycles rather than at the termination of cycles. Thus, decision

makers can improve processes while they are in progress. Establishing a progress monitoring system at the beginning of the program will help evaluators and decision makers work together to develop a clear understanding of program processes. The application of management information tools such as PERT and Critical Path Method (CPM) are useful for this task, especially if the program is in the initial stages of implementation.

Because program evaluations must delineate, obtain, and disseminate information at various levels of a program, multiple methods are more likely to be successful than a single approach. The PDK Committee lists several methods for obtaining and delineating information: questionnaires, hearings, Delphi Technique, survey committees, board meetings, and staff advising groups. Suggestions for disseminating information include published reports, lecture presentations, panel presentations, videotapes, individual presentations, and media reports. Such methods are quite different from those used in experimental research.

In summary, the purpose and methodologies involved in program evaluation are usually different from that of research even though the two endeavors may occur simultaneously within the same program. Research seeks to add new knowledge to extant knowledge bases that generalize to wide populations, while program evaluation seeks to describe and improve program components within unique environments. The PDK Committee points out that decision rules for conventional experimental research and the inferential tests used to detect significant differences between treatments are binary in nature and lead to go no-go; reject, not reject; better, not better conclusions. Such rules are sometimes appropriate for making summative conclusions about a program component or even an entire program, but they are inadequate for formative program evaluation.

CONCLUSIONS

Summary of the Evolution of Program Evaluation

The birth of many large educational programs in the 1960s was followed by the passing into law of ESEA, the Congressional Budget and Improvement Control Act, and the fostering of a general initiative for greater accountability in Government-funded programs. This growth brought about a substantial increase in the need for program evaluation. However, early evaluation efforts failed to (a) measure the degree of program success and (b) provide program decision makers with information to improve their programs. This crisis led to the development of new organizations devoted to clarifying the purposes and processes involved in program evaluation and developing new evaluation procedures. Many new journals and books were published during the late 1960s and early 1970s that focused on program evaluation. A substantial number of these publications are relevant to the philosophy, theory, and methodology of educational program evaluation. Since the 1960s, evaluation

researchers have expanded the concept and definition of program evaluation and have assembled a substantial number of new methods, tools, and strategies for practitioners. Many of the shortcomings of recent program evaluations appear to be due to (a) the failure of evaluators to investigate and heed the lessons learned during past evaluation efforts, (b) the lack of or misapplied use of evaluation tools, and (c) the failure of program sponsors, decision makers, and evaluators to develop a clear understanding of evaluation's roles within program planning, implementation, and operation.

Information Sources for Evaluation Practitioners and Researchers

Although funds for program evaluation and efforts to advance the field appear to have dwindled in the last 15 years, many periodicals and books that describe case studies, review new evaluation methods and tools, and present theoretical perspectives on program evaluation are available today. Government publications describe case studies of evaluations, methods employed, and conclusions made about specific programs. Journal articles focus on the advocation of new methods, development of theory, and recommendations for practitioners. Each article usually deals with a specific topic. More comprehensive coverage of the field can be obtained from books about program evaluation and related disciplines (e.g., policy analysis). Because of the great complexity and costs associated with evaluating programs, it is strongly recommended that prospective evaluators become well versed in the literature prior to participating in planning a new program or attempting to evaluate an existing program.

Lessons Learned in Program Evaluation

The investigation of program evaluation and subsequent literature published by evaluation researchers has produced several lessons for practitioners and researchers:

(a) Evaluators and program decision makers need to share the same understanding of the purpose and expected outcomes of an evaluation.

(b) Evaluators need to have the required skills for conducting several different activities that make up well-planned evaluations.

(c) Guidelines that have been developed for each type of evaluation activity should be thoroughly studied by practitioners even before planning the evaluation.

(d) Evaluation tools and methods should be exploited by evaluators to ensure a reduction in the cost of evaluation processes, a high quality of data collection and analysis, the effective dissemination of information, and useful documentation.

Many serious problems can be avoided if appropriate planning for evaluation can be accomplished with the above lessons in mind.

The most promising predictor of success of program evaluation appears to be proper planning. Planning for evaluation should take place during the earliest stages of a new program. The goals and procedures of an evaluation should be expected to evolve and expand as the program matures from the planning stage, to its implementation, and finally to its fully operational status. During the planning stage, evaluators should work closely with program decision makers and sponsors to ensure that evaluation goals, objectives, and outcomes are understood and agreed upon. As the program is implemented, evaluators must provide information to decision makers that will maximize the continuity and efficiency of program components. Formative evaluation should be conducted on each life cycle of the program, and results and conclusions should be provided to decision makers within predesignated time limits so they can utilize the information prior to the following cycle. Computer-assisted evaluation tools are a must for this task. Summative measures should take place only after the program has become fully operational within its designated context and is functioning at its intended capacity. A formative evaluation component should function throughout the life of the program and continue to provide information that will drive future improvements and ensure the efficiency of new program components as they evolve.

REFERENCES

Alemi, F. (1987). Subjective and objective methods of evaluating social programs *Evaluation Review, 11*(6), 765-774.

Boruch, R. F. (1975). On common contentions about randomized field experiments. In R. F. Boruch, & H. W. Riecken (Eds.), *Experimental testing of public policy: The proceedings of the 1974 Social Science Research Council Conference on Social Experiments* (pp. 107-142) Boulder, CO: Westview Press.

Boruch, R. F., Cordray, D. S., Pion, G. M., & Leviton, L. C. (1983). Recommendations to Congress and their rationale: The Holtman Project. *Evaluation Review, 7*(1), 5-35.

Brandt, R. (1978). On evaluation: An interview with Daniel L. Stufflebeam. *Educational Leadership, 35*(4), 249-254.

Braybrooke, D., & Lindblom, C. E. (1963). *A strategy of decision*. New York, NY: The Free Press.

Britan, G. M. (1978). Experimental and contextual models of program evaluation. *Evaluation and Program Planning, 1*(3), 229-234.

Bruce, P. D., Killion, T. H., & Rockway, M. R. (1989). *Aircrew training evaluation: B-52 and KC-135 formal school training* (AFHRL-TR-88-49, AD A208 860). Williams AFB, AZ: Operations Training Division.

Chelimsky, E. (1985). Comparing and contrasting auditing and evaluation. *Evaluation Review, 9*(4), 483-503.

Ciechinelli, L. T., Harmon, K. R., & Keller, R. A. (1982). *Relative cost and training effectiveness of the 6883 F-111 converter/flight control system simulator as compared to actual equipment* (AFHRL-TR-82-30, AD-A123 534). Lowry AFB, CO: Logistics and Technical Training Division.

Cohen, A. B., Hall, K. C., & Cohodes, D. R. (1985). Evaluation readiness. *Evaluation and Program Planning, 8*(4), 315-326.

Comptroller General of the U. S. (1979). *The Air Force should cancel plans to acquire two computer systems at most bases* (FGMSD-80-15). General Accounting Office.

Cook, T. D., & Gruder, C. L. (1978). Meta-evaluation research. *Evaluation Quarterly, 2*(1), 5-51.

Cook, D. L., & Leviton, L. C. (1983). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Evaluation Studies Review Annual, 8*(2), 59-82.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornick, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass, Inc.

Dobson, D., & Cook, T. J. (1979). Implementing random assignment: A computer-based approach in a field experimental setting. *Evaluation Quarterly, 3*(3), 472-489.

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Evaluation Research Society Standards Committee. (1982). Evaluation research society standards for program evaluation. In *Standards for evaluation practice: New directions for program evaluation*, No. 15 (pp. 7-19). San Francisco, CA: Jossey-Bass, Inc.

Filstead, W. J. (1981). Using qualitative methods in evaluation research: An illustrative bibliography. *Evaluation Review, 5*(2), 259-268.

Gagné, R. M., & Briggs, L. J. (1979). *Principles of instructional design* (2nd ed.). New York, NY: Holt, Rinehart, & Winston.

Greene, J., & McClintock, C. (1985). *Triangulation in evaluation: Design and analysis issues. Evaluation Review, 9*(5), 523-545.

Haskins, J. B. (1981). A precise notational system for planning and analysis. *Evaluation Review, 5*(1), 33-50.

Hays, J. S. (1987). Evaluation and management in local human service agencies: An agenda for collaboration. In J. S. Wholey (Ed.), *Organizational excellence* (pp. 31-44). Lexington, MA: D. C. Heath and Company.

Health and Human Services Evaluation and Documentation Center. (1985). *Compendium of Health and Human Services (HHS) evaluations and relevant other studies*. HHS Evaluation and Documentation Center, Office of Assistant Secretary for Planning and Evaluation.

Hedges, L. V. (1983). Estimation of effect size from a series of independent experiments. *Evaluation Studies Review Annual, 8*(11), 205-214.

Hedges, L. V. (1986). Advances in statistical methods for meta-analysis. *Evaluation Studies Review Annual, 11*(40), 731-748.

Heilman, J. G. (1980). Paradigmatic choices in evaluation methodology. *Evaluation Review*, *4*(5), 693-712.

Horn, W. F., & Heerboth, J. (1982). Single-case experimental designs and program evaluation. *Evaluation Review*, *6*(3), 403-424.

House, E. R. (1986). Internal evaluation. *Evaluation Practice*, *7*(1), 63-64.

Jemelka, R. P., & Borich, G. D. (1979). Traditional and emerging definitions of educational evaluation. *Evaluation Review*, *3*(2), 263-276.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials*. New York, NY: McGraw-Hill.

Kennedy, M. M. (1979). Generalizing from single case studies. *Evaluation Quarterly*, *3*(4), 661-678.

Leithwood, K. A., & Montgomery, D. J. (1980). Evaluating program implementation. *Evaluation Review*, *4*(2), 193-214.

McClintock, C. (1984). Toward a theory of formative program evaluation. In D. Deshler (Ed.), *Evaluations for program improvement: New directions for program evaluation*, no. 24 (pp. 77-95). San Francisco, CA: Jossey-Bass, Inc.

Nagel, S. S. (1985). Evaluation research and policy studies. *Evaluation News*, *6*, 59-64.

Nagel, S. S. (1986). An overview of microcomputers and evaluation research. *Evaluation Review*, *10*(5), 563-577.

Nay, J. N., & Kay, P. (1982). *Government oversight and evaluability assessment*. Lexington, MA: D.C. Heath and Company.

Nielsen, L., & Turner, S. D. (1983). Program evaluation as an evolutionary process. *Evaluation Review*, *7*(3), 397-405.

Petty, R. E., & Cacioppo, J. T. (1981). *Attitudes and persuasion: Classic and contemporary approaches*. Dubuque, IA: Wm. C. Brown Company Publishers.

Pollard, W. E., Cooper, A. C., & Griffin, D. H. (1985). Documentation in evaluation research: Management and scientific requirements. *Evaluation and Program Planning*, *8*(2), 161-169.

Quade, E. S. (Ed.). (1967). *Analysis for military decisions*. Chicago, IL: Rand McNally & Company.

Reigeluth, C. M., & Merrill, D. M. (1984). *Extended task analysis procedure (ETAP)*. Lanham, MD: University Press of America, Inc. 96 pages.

Reineke, R. A., & Welch, W. W. (1986). Client-centered meta-evaluation. *Evaluation Practice*, *7*(3), 16-34.

Roberts-Grey, C., Buller, A., & Sparkman, A. (1987). Linking data with action: Procedures for developing recommendations. *Evaluation Review*, *11*(5), 678-684.

Roos, L. L., Jr., Nicol, J. P., Johnson, C. F., & Roos, N. P. (1979). Using administrative data banks for research and evaluation. *Evaluation Quarterly*, *3*(2), 236-255.

Ross, J. A. (1980). Decision rules in program evaluation. *Evaluation Review*, *4*(1), 59-74.

Rossi, R. J., & McLaughlin, D. H. (1979). Establishing evaluation objectives. *Evaluation Quarterly*, *3*(3), 331-346.

Rossman, G. B., & Wilson, B. L. (1985). Numbers and words: Combining quantitative and qualitative methods in a single large-scale evaluation study. *Evaluation Review*, *9*(5), 627-643.

Rutman, L. (1980). *Planning useful evaluations*. Beverly Hills, CA: Sage Publications.

Scheirer, M. A., & Rezmovic, E. L. (1983). Measuring the degree of program implementation. *Evaluation Review*, *7*(5), 599-634.

Scriven, M. (1967). *The methodology of evaluation* (AERA Monograph Series on Curriculum Evaluation). Chicago, IL: Rand McNally & Company.

Scriven, M. (1981). *Evaluation Thesaurus* (3rd ed.). Pt. Reyes, CA: Edge Press.

Scriven, M. (1986). New frontiers of evaluation. *Evaluation Practice*, *7*(1), 7-44.

Shadish, Jr., W. R., & Reichardt, C. S. (1988). Journals that publish work of interest to evaluators. *Evaluation Practice*, *9*(3), 29-31.

Sherrill, S. (1984). Toward a coherent view of evaluation. *Evaluation Review*, *8*(4), 443-466.

Siegel, K., & Tuckel, P. (1985). The utilization of evaluation research: A case analysis. *Evaluation Review*, *9*(3), 307-328.

Stake, R. E. (1967). The countenance of educational evaluation. *Teachers College Record, 68*(7), 523-540.

Strosberg, M. A., & Wholey, J. S. (1983). Evaluability assessment: From theory to practice in the Department of Health and Human Services. *Public Administration Review, 43*, 66-71.

Stufflebeam, D. L. (1975). *Meta-evaluation*. Occasional Paper #3, Evaluation Center, Western Michigan University.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. M. (1971). *Educational evaluation and decision making*. Bloomington, IN: Phi Delta Kappa, Inc.

Stufflebeam, D. L., & Shinkfield, A. J. (1985). *Systematic evaluation*. Boston, MA: Kluwer-Nijhoff.

Stufflebeam, D. L., & Webster, W. J. (1981). An analysis of alternative approaches to evaluation. *Evaluation Studies Review Annual, 6*(4), 70-85.

Suchman, E. A. (1967). *Evaluative research*. New York, NY: Russell Sage Foundation.

Superintendent of Documents. (1978-1989). *Monthly catalogue of Government publications*. Washington, DC: U.S. Government Printing Office.

Thompson, M. S. (1982). *Decision analysis for program evaluation*. Cambridge, MA: Ballinger Publishing Company.

Thorndike, R. L., & Hagen, E. (1961). *Measurement and evaluation in psychology and education*. New York, NY: John Wiley & Sons, Inc.

Turpin, R. S., Smith, N. L., & Darcy, L. A. (1987). Survey of journals publishing evaluation research. *Evaluation Practice, 8*(2), 10-19.

Tyler, R. W. (1967). Changing concepts of educational evaluation. In R. E. Stake (ed.), *Perspectives of curriculum evaluation* (Vol. 1). New York: Rand McNally.

U.S. Army Research Institute for the Behavioral and Social Sciences. (1978). Review of peer evaluation research.

U.S. General Accounting Office. (1988). *Program evaluation issues* (GAO Publication No. OCG/-89-8TR). Washington, D.C.: U.S. General Accounting Office.

Votey, H. L., Jr. (1981). Constructing a quasi-controlled environment. *Evaluation Review, 5*(1), 90-122.

Wholey, J. S. (Ed.) (1987). *Organizational excellence*. Lexington, MA: D.C. Heath and Company.

Worthen, B. R., & Sanders, J. R. (1987). *Educational evaluation*. West Plains, NY: Longman, Inc.

*U.S. GOVERNMENT PRINTING OFFICE 1992-0-701-052/60044